Early Fusion of Traditional and Deep Features for Irony Detection in Twitter

Hairo Ulises Miranda-Belmonte[1] and Adrián Pastor López-Monroy[2]

[1] Mathematics Reseach Center (CIMAT) Campus Monterrey,
Alianza Centro 502, 66629 Nuevo León, México
hairo.miranda@cimat.mx
[2] Mathematics Research Center (CIMAT),
Jalisco s/n Valenciana, 36023 Guanajuato, México
pastor.lopez@cimat.mx

Abstract. In this paper we describe the system designed by the Mathematics Research Center (CIMAT) for participating at *IroSvA: Irony Detection in Spanish Variants 2019*. In this work, we addressed the Irony Detection task by exploiting Traditional and Deep Features. The core idea is to separately extract traditional features based on n-gram occurrences and frequencies, and jointly use them with features computed from neural networks. For this, we concatenate three different feature vectors that feed a classifier (a.k.a., *early fusion*), each attribute space captures information from: i) n-grams, ii) embeddings and iii) hidden units from a Recurrent Neural Network. This strategy produces enriched representations for documents, where different aspects of style and content are highlighted for the classifier. We evaluate the proposed approach and compare with the individual performance of each feature space. The experimental evaluation on the IroSvA corpora showed that the proposal outperforms all baselines, has the best performance in the Cuban variant of Spanish and the overall second place in the challenge. Furthermore, the results are strong evidence of the usefulness of the representation to identify irony independently of the language variety.

Keywords: Irony Detection · Text Classification · Embeddings for Classification

1 Introduction

Recently, the Irony Detection task has gained the interest of the scientific community. In Irony Detection one aims to build computational methods that automatically identify this phenomenon in written language [3, 16, 6]. The task is very challenging and has a wide applicability with a broad impact in a number of problems ranging from marketing to business intelligence. The Irony Detection

task at IroSvA 2019 is focused on the recognition of this phenomenon in three different variants of Spanish: Cuba, Spain and Mexico [10]. According to the literature, the Irony Detection task has been approached by several researchers [14, 15]. Many of these efforts have been devoted to the analysis of the textual representation and features (e.g., lexical, syntactical, etc.) [4, 3, 16]. Regardless of the novel textual features and representations, most of them fail in capturing accurate information from informal documents and solve the problem [3, 15]. This is especially true in the social media domain, where the easiness of writing/sending messages leads people to make many grammatical and spelling errors. The previous situation captured the attention of some researchers, who began to model high level aspects (e.g., semantic and structural information) for the Irony Detection task [15, 10].

According to the literature in text classification in social media, some methods that build high level features (e.g., embeddings, LSTMs, CNNs, etc.) based on different textual units (e.g., word or character level) have been useful for boosting the performance [12, 13, 10]. In this paper we propose to study these features together with traditional features based on word and n-gram statistics. We propose to extract finer and very local patterns by using n-grams and more global patterns by using LSTMs and embeddings. For this we propose the idea of performing early fusion over the latter feature spaces. The idea is to separately compute each feature set and then perform the concatenation that will feed a Support Vector Machine (SVM). Experimental results and a exhaustive evaluation using the latter ideas seem promising and competitive compared to other approaches that uses individually each feature space.

The remainder of this paper is organized as follows: Section 2 introduces the proposed approach. Section 3 explains the evaluation methodology for this proposal. Finally Section 4 outlines the main conclusions and future avenues of inquiry.

2 Early Fusion of Traditional and Deep Features

In this work we aim to jointly exploit Traditional and Deep Features. The traditional attributes will be word and n-gram statistics, whereas the deep features will be those computed by means of LSTMs and embeddings (Word2Vec). We detail each of the three feature spaces in the following lines:

1. The Bag of Words (BoW) is the most standard strategy to extract and use traditional features based on counting words. In this paper we refer to this model as Bag of Terms (BoT) as the generalization that could have also other types of features such as n-grams at character and word level. In our methodology, we always use the well known Term Frequency Inverse Document Frequency (TF-IDF) weighting scheme for this feature space.
2. The embeddings are word vectors learned by using Word2Vec from big text collections. In this work we evaluated pretrained word embeddings of 300 dimensions from Google [9] and 400 dimensions from Twitter [5]. The idea is

to represent the documents by averaging the vectors of words that each document contains. The intuition is to keep the linguistic regularities contained in each document.
3. The Long Short Term Memory Neural Networks (LSTMs) are also used in order to capture long term dependencies in documents. We evaluate LSTM in two ways, in the first one we encode the whole tweet with the LSTM, then we use the hidden layer in the last time step to fed a Support Vector Machine (SVM). In the second strategy we simply use the sigmoid function to classify with the LSTM in the last step.

In the following section we will present a detailed evaluation of the previous strategies and different combinations of them. Each outcome in the results is a particular setting for each feature space. For example the BoT uses different number of unigrams according to a minimum frequency in documents in each collection. In particular we discarded all terms that occur in less of the 10% of documents. Moreover, when specified, the BoT uses n-grams (e.g., BoT-n-grams), that means the use of different number of features that we automatically selected by using the chi square strategy implemented in sklearn. For the case of embeddings we use the pretrained word vectors from [9] and [5]. Finally, for the LSTM we used 256 hidden units, 50 bach size and Adam optimizer.

3 Evaluation

In these experiments we are interested in exploring the contribution of different strategies to classify in the Irony Detection task. The target scenarios are language varieties. For this we present results from several methods and the combination of them.

Table 1. Experimental evaluation for Irony Detection. Results show the average F-1 performance of different representations that are fed into a SVM.

Macro F-1 score in IroSvA corpora			
Representation	Mexico	Spain	Cuba
BoT	**64.27**	**68.50**	60.10
W2V-Google	62.53	66.00	60.64
W2V-Twitter	62.60	65.25	60.65
LSTM	63.30	66.75	**60.85**

In Tables 1 and 2 we evaluate the performance of different representations using a Support Vector Machine (SVM) and Neural Network (NN) respectively. The SVM is the one implemented in sklearn (LinearSVC), and we only optimize the C parameter, whereas the NN was implemented in Keras and had only one hidden layer with 512 neurons, dropout of .25, batch size of 50, nadam optimizer and used 20% of the training data for validation. From these results it can be

Table 2. Experimental evaluation for Irony Detection. Results show the average F-1 performance of different representations that are fed into a Neural Network.

Macro F-1 score in IroSvA corpora			
Representation	Mexico	Spain	Cuba
BoT	**60.47**	**67.31**	**57.90**
W2V-Google	59.75	61.35	55.57
W2V-Twitter	55.65	62.36	52.82
LSTM	60.21	63.25	56.41

seen that the SVM, in general, outperforms the Neural Network by a considerable difference. Also note that the Bag of Terms (BoT) using only unigrams (words) has the overall best performance. This is interesting, since the BoT has shown outstanding performance in many different text classification tasks [2, 4, 10, 8, 1, 11, 7]. Also note that, Word2Vec (W2V) showed better performance when trained on Google instead of Twitter. Furthermore, the LSTM showed a sightly better performance than embeddings. We hypothesize that this is because of long term dependencies captured by this model and a finer-level of granularity in the representation. Finally Word2Vec-Twitter representation obtained the overall lowest performance, this could be due to the few training documents in the data collection, where more data could be necessary to have more words in the pre-computed vocabulary.

Table 3 shows the final results of this evaluation. The table shows the performance of the combined representations. According to these results, the best performing strategy is the early fusion of the three feature spaces. This is the traditional BoT, Google embeddings [9] and the hidden units of the LSTM. Note that in this table we now show an additional result that extends BoT with $1-5$-grams at word and character level. In this experiment we empirically select the best n-grams for each size of n according to the Chi square metric. The particular setting for BoT(n-grams) is as follows: i) 5k unigrams, 3k bigrams and 1k trigrams at word level, and ii) 5k 3-grams, 5k 4-grams and 5k 5-grams at character level. The idea of this is to give much more information to the BoT. The latter is shown in the last row of the table and has the overall best performance for our experiments. We believe that the good performance in the latter approach is because finding patterns among different types of attributes, provides a more detailed perspective for documents. Moreover, the embeddings and LSTMs help to capture more information in a global level of the documents. In this regard, the proposal is a suitable representation in Irony Detection task for social media. Thus, the approach presented in this paper is an effective alternative to address the Irony Detection task in different language variants, where documents present challenging difficulties that hinder the accurate work of most natural language processing tools.

Table 3. Experimental evaluation for Irony Detection. Results show the average F-1 performance of different combined representations that are fed into a SVM.

Macro F-1 score in IroSvA corpora			
Representation	Mexico	Spain	Cuba
BoT+W2V-Google	65.40	**69.77**	61.88
BoT+W2V-Twitter	63.58	67.68	62.15
BoT+LSTM	64.57	68.66	61.81
BoT+W2V-Google+LSTM	66.12	69.04	62.33
BoT	64.27	68.50	60.10
BoT(n-grams)	65.01	68.79	62.05
BoT(n-grams)+W2V-Google+LSTM	**67.09**	64.49	**65.96**

4 Conclusions

In this paper we presented a novel idea to approach the Irony detection task. The proposal extracts traditional and deep features from user-documents in the dataset. For this, we exploit the n-grams, embeddings and the hidden layer of a LSTM. Such features model target user-documents by performing early fusion of features that are fed into a SVM. The intuitive idea is to encode local patterns with n-grams and embeddings, and longer dependencies between words with the LSTM. The latter strategy helps to improve the classification performance in most of the language varieties. Using all these attributes, the classifier can keep good classification rates. This is due to the relationship among different kind of attributes. We have shown better experimental results than the standard BoT, Word2Vec and LSTM, which has shown to be useful, but better when jointly used.

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Meza, I.: Evaluating topic-based representations for author profiling in social media. In: Ibero-American Conference on Artificial Intelligence. pp. 151–162. Springer (2016)
2. Bosco, C., Patti, V., Bolioli, A.: Developing corpora for sentiment analysis: The case of irony and senti-tut. IEEE Intelligent Systems **28**(2), 55–63 (2013)
3. Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in twitter and amazon. In: Proceedings of the fourteenth conference on computational natural language learning. pp. 107–116. Association for Computational Linguistics (2010)
4. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 470–478 (2015)
5. Godin, F., Vandersmissen, B., De Neve, W., Van de Walle, R.: Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using

distributed word representations. In: Proceedings of the Workshop on Noisy User-generated Text. pp. 146–153 (2015)
6. González-Ibánez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2. pp. 581–586. Association for Computational Linguistics (2011)
7. López-Monroy, A.P., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Stamatatos, E.: Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems **89**, 134 – 147 (2015)
8. López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A new document author representation for authorship attribution. In: Mexican Conference on Pattern Recognition. pp. 283–292. Springer (2012)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
10. Ortega-Bueno, R., Rangel, F., Hernández Farías, D.I., Rosso, P., Montes-y-Gómez, M., Medina Pagola, J.E.: Overview of the Task on Irony Detection in Spanish Variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS.org (2019)
11. Ortega-Mendoza, R.M., Franco-Arcega, A., López-Monroy, A.P., Montes-y Gómez, M.: I, me, mine: The role of personal phrases in author profiling. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 110–122. Springer (2016)
12. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF. sn (2015)
13. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF (2016)
14. Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowledge and Information Systems **40**(3), 595–614 (2014)
15. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Language resources and evaluation **47**(1), 239–268 (2013)
16. Veale, T., Hao, Y.: Detecting ironic intent in creative comparisons. In: ECAI. vol. 215, pp. 765–770 (2010)